Finite Mixture Models

Dias, J.G.

Summary

Finite mixture (FM) models have become a standard tool to address research questions in different scientific fields. This thesis provides an overview of the methodology, some applications, and further extensions of the theory. Chapter 2 gives an overview of the field and provides the setting for Chapters 3 to 6.

Chapter 3 illustrates FM modelling with the purpose of improving the effectiveness of the social marketing of family planning programmes. It is shown that the use of *post hoc* segmentation techniques could effectively identify subpopulations that have similar intentions and behaviour and are likely to respond similarly to family planning programmes.

Chapter 4 explores the use of bootstrap methods in FM models. It shows that these methods can help measure both parameter and classification uncertainties. It is also shown that the exact implementation of this computer-intensive technique for FM models is not straightforward. In particular, the label-switching problem can be severe and has to be addressed with care. Measures of classification uncertainty at aggregate and individual levels were discussed. Aggregate or bootstrap classifiers for model-based clustering are proposed as an alternative to the Bayes classifier.

Chapter 5 focuses on model selection based on the information criteria for FM models. A large-scale Monte Carlo study comparing 17 information criteria is performed. A new procedure is introduced to control the level of separation of components. It is found that the Akaike Information criterion with penalisation factor of 3 (AIC3) has the best overall performance of all the information criteria.

Chapter 6 compares the EM, SEM and MCMC algorithms for univariate Gaussian mixture models. Their relative performance for log-likelihoods with particularly problematic shapes, involving ridges and/or saddle points, is investigated. We found that the convergence of the EM algorithm can be extremely slow and dependent upon the type of starting values, stopping rule, and tolerance level. As a consequence, the EM algorithm often fails to converge to the global maximum of the log-likelihood surface. It is shown that the SEM and MCMC algorithms escape from saddle points in the log-likelihood as a result of the simulation steps. The effect of different prior specifications and label-switching strategies are discussed as well.